SYNOPSIS OF RESEARCH PROJECT

1.	**Name of the Student**	**:**	CHOVATIYA PRUTHVI VALLABHBHAI
2.	**Registration No.**	**:**	1010125001
3.	**Faculty**	**:**	Agriculture
4.	**Name of the Degree Submitted for**	**:**	Ph. D.
5.	**Major Subject**	**:**	Agricultural Statistics
6.	**Minor Subject**	**:**	Agricultural Economics
7.	**Name of Major Guide**	**:**	Dr. M. S. Shitap Assistant Professor Department of Agricultural Statistics College of Agriculture JAU, Junagadh.
8.	**Title of Research Project**	**:**	**PREDICTION OF RAINFALL USING ADVANCED DATA-DRIVEN MACHINE LEARNING TECHNIQUES IN SAURASHTRA REGION**

9. Introduction

Agriculture, being one of the oldest and most critical human activities, is increasingly relying on advanced computational methods for better prediction and decision-making. In India, approximately 60–70% of the total population directly or indirectly depends on agriculture and its allied sectors, which contribute 17.94% to the national Gross Value Added (GVA) at current prices in year of 2024-25 (Anonymous, 2025). Forecasting is a fundamental analytical tool used to predict future trends, patterns and behaviors in complex data systems. It finds extensive application across diverse fields including weather forecasting, economic modeling, energy management, transportation planning, sales analytics and agricultural management. Given this immense socioeconomic significance, accurate prediction of agricultural outcomes such as climatic conditions like precipitation is essential for sustainable agricultural planning and food security.

India's diverse topography makes it possible to grow a variety of goods, including cash crops, fruits, vegetables, pulses and cereals. Nonetheless, India's agricultural industry is still reliant on monsoon rainfall, particularly during the south-

west monsoon. The months of June, July, August and September receive more than 90% of the country's yearly rainfall during the monsoon season in India.

Precipitation is one of the most important climatic factors influencing agricultural productivity. Precipitation in particular has a significant impact on agricultural productivity. Crop scheduling, irrigation needs and soil moisture availability are all influenced by precipitation. Accurate precipitation prediction has become essential for farmers, hydrologists, agricultural planners and disaster management authorities due to the rising frequency of extreme weather events brought on by global climate change.

Kutch, North Gujarat, East Central, Saurashtra and South Gujarat are the five zones that make up the state of Gujarat. Between 1992 and 2021, the average rainfall in these zones was 456 mm, 720 mm, 806 mm, 717 mm and 1476 mm, respectively (Anonymous, 2022). Gujarat is one of the agriculture dominant areas where rainfall plays a very important role in the agriculture process and functions. Due to water scarcity, cultivation during non-rainy months became difficult and water became an essential input (Lodhavi, 2022). Global climate change has recently changed rainfall patterns, increasing the frequency and severity of extreme precipitation events. As a result, precise precipitation forecasting is crucial to controlling the effects of these occurrences on infrastructure and populations. Developing reliable and accurate precipitation forecasting models is essential to mitigating the impact of extreme events (Schumann *et al.*, 2016; Mehta and Yadav, 2021).

The presence and severity of rainfall are influenced either directly or indirectly by a number of variables, including temperature, relative humidity, sunshine, pressure, evaporation, etc. To comprehend the climatic pattern for a certain area, it is crucial to analyze patterns in meteorological parameters. Crop output is significantly impacted by rainfall and other meteorological factors. Farmers and agricultural planners make decisions to maximize crop patterns by analyzing past data and forecasting future trends. Generally, weather parameter trend is found out using non parametric method like Mann Kendall. In India, many researchers Kumar *et al.,* (2017) and Rajput *et al.,* (2025) used the Mann Kendall test to find out the trend of weather parameter in India as well as Gujarat.

For prediction tasks, conventional statistical models like Multiple Linear Regression (MLR) have long been employed. Nevertheless, the intricate and nonlinear relationships found in actual agricultural and meteorological datasets are frequently

missed by these models. Machine learning (ML) and artificial intelligence (AI) based models, such as Multivariate Adaptive Regression Splines (MARS), Artificial Neural Network (ANN), Support Vector Regression (SVR), Random Forest (RF) and Extreme Gradient Boosting (XGBoost), have become potent substitutes with superior predictive accuracy and flexibility due to the quick development of computational technology.

Even though each of these models has its own advantages, no single model is the best at all datasets and prediction tasks. Recent prediction literature has focused a lot of attention on hybrid modeling approaches that combine the complementary strengths of two or more individual models. MARS is especially good at identifying significant input variables and capturing intricate nonlinear interactions because it is a naturally interpretable and nonparametric model. MARS-based hybrid models may provide better predictive performance when paired with ML algorithms such as ANN and SVR. This is because the coupled ML model manages the residual complex patterns, while MARS is used for feature selection and nonlinear basis function generation. The present study is motivated by the need to develop, compare and validate such standalone and hybrid forecasting models for daily or weekly precipitation.

1.2 Predicting Models: An Overview

Statistical models and artificial intelligence (AI)/machine learning models are two broad categories of prediction models. MLR and other statistical models are based on certain distributional characteristics of data and are based on probability theory. Although MLR is straightforward and easy to understand, it makes the assumption that variables are linear, which is frequently broken in intricate agricultural and meteorological datasets.

Friedman (1991) created MARS, a nonparametric regression method that uses piecewise linear basis functions to automatically model nonlinearities and interactions between predictor variables. It is especially helpful for datasets with a lot of dimensions and can be used as both a predictive model and a tool for choosing which variables to include. Various machine learning techniques to increase the precision of rainfall forecasts on a daily, monthly and annual basis. It has been demonstrated that machine learning methods are superior to the conventional deterministic approach for rainfall prediction. ANN, which is based on biological neural networks, is great at learning how to map complex nonlinear relationships. SVR applies support vector machine principles to regression tasks, offering robustness against outliers and good generalization. Random Forest is a group of decision trees that uses bootstrap aggregation and random

feature selection to avoid overfitting. This makes it very accurate for tabular data, where decision trees are models that divide the feature space recursively. XGBoost, based on gradient boosting, sequentially builds trees to minimize prediction error, representing one of the most powerful ensembles learning algorithms for structured data.

Hybrid models combine the feature selection and nonlinear modeling capabilities of MARS with the learning power of ML algorithms. The MARS-ANN hybrid takes inputs from MARS-derived basis functions and sends them to ANN. MARS-SVR uses MARS to create features and then SVR to make predictions. In a number of studies on agricultural and weather prediction, these kinds of hybrid frameworks would show better accuracy than single models.

The Stochastic Markov Chain model is the most successful prediction model for long-term rainfall patterns. This model is helpful for expressing the temporal dependency and variability of rainfall occurrences since it is based on probabilistic transitions between several rainfall states or categories. Based on past transition probabilities, this model can simulate the incidence and intensity of rainfall. Many scholars in India have employed the Markov chain model to comprehend or forecast rainfall Chowdhury *et al.*, (2025) and Goel and Singh (2025).

10. Practical Utility of the Research Project

The accurate prediction of agricultural variables has become an essential objective in modern agricultural science, owing to the increasing demands on food systems to satisfy the needs of a rapidly expanding global population amidst rising climate variability and resource limitations. The Saurashtra region of Gujarat, characterized by its semi-arid agro-climatic conditions and near-complete dependence on monsoon rainfall for agricultural sustenance, presents a particularly challenging yet critically important case study for precipitation forecasting research. meteorological data shows complex structural characteristics that make precise modelling extremely difficult. Such data often show both linear and nonlinear autocorrelation patterns at the same time. They are also nonstationary, very volatile, have irregular seasonal changes and have sudden structural breaks caused by climate shocks, policy changes and market disruptions. These inherent complexities make conventional linear statistical models, such as Multiple Linear Regression, insufficient as independent prediction instruments, as they are fundamentally limited by assumptions of linearity, stationarity and constant variation that are frequently violated in actual agricultural datasets. The progress of machine learning and hybrid modelling frameworks has given researchers a wider range

of tools that can capture both the deterministic trend components and the stochastic nonlinear patterns at the same time. Models like Multivariate Adaptive Regression Splines, Support Vector Regression, Artificial Neural Networks, Random Forest and XGBoost have all shown promising prediction abilities in different agricultural contexts. However, empirical evidence shows that no one specific model works best for all types of data, commodities and time frames. This has led to the creation of hybrid frameworks that strategically combine the strengths of several modelling approaches. A significant amount of literature will look at the forecasting of precipitation as the most important factors influencing agricultural productivity and planning within this developing field of agricultural prediction research. With the dual goal of increasing the prediction accuracy of precipitation, the proposed research will build upon this rich body of literature by developing and evaluating a comprehensive hybrid predicting framework that integrates MARS as a preprocessing and feature selection stage with multiple machine learning algorithms, such as ANN and SVR. Second, the study undertakes a rigorous prior analysis of temporal trends in key meteorological parameters across the Saurashtra region using the Mann-Kendall non-parametric trend test. Third, the temporal variation and distributional characteristics of precipitation across the Saurashtra region are systematically examined to provide a foundational understanding of the rainfall regime. Also, Markov Chain probability approach is further employed to model the stochastic occurrence of wet and dry spells, offering probabilistic forecasting capabilities that are particularly valuable for agricultural scheduling purposes. Thus, in the area of climate-resilient agriculture, methodologically sound and practically implementable forecasting tools can significantly assist farmer decision-making, agricultural extension services and evidence-based policymaking.

11. Objectives

1. To understand variations in saurashtra region precipitation
2. To examine the trend in various meteorological parameters of saurashtra region
3. To use machine learning techniques to predict both short and long term precipitation
4. To compare the efficiency of developed models

12. Reviews of Literature

12.1 Trend analysis of weather parameter

Mondal *et al.* (2012) discovered a rainfall trend in the northeastern region of Orissa state's Cuttack district. The monthly variability of rainfall was determined for this study using daily rainfall data over 40 years, from 1971 to 2010. The outcome indicated that the Cuttack district had generally minor adjustments.

Kumar *et al.* (2017) examined the rainfall pattern in order to identify long-term trends in the South Gujarat region. The study used 115 years of monthly and yearly rainfall data for five South Gujarati locations for Navsari, Bharuch, Surat, Valsad and Dang from period of 1901 to 2015. The findings showed that while annual and monsoon rainfall was dropping in Navsari and Bharuch districts, it was increasing in Valsad, Dang and Surat.

Mehta *et al.* (2022) investigated rainfall variability in the Surat area of Gujarat using mean monthly and seasonal rainfall from 1981 to 2000. The findings showed a substantial negative trend in the annual rainfall data.

Joshi *et al.* (2023) examined the yearly and monthly trends in the maximum and lowest temperatures in Bhogavo River Watersheds (BRW) in western India from 1979 to 2014. In addition to observing the trend, the study assessed extreme temperatures. The yearly minimum and maximum air temperature showed a nonsignificant trend, according to the results.

12.2 Time series analysis

Raha and Gayen (2020) analyzed meteorological drought in West Bengal's Bankura District using exponential smoothing models and daily data gathered from the Climate Forecast System Reanalysis (CFSR) in SWAT format over a 35-year period (1979–2014). Holt Winter Exponential Smoothing (HWES) and Double Exponential (DE) were used to examine the study over multiple time steps (3 months, 6 months, 12 months, 24 months and 48 months). The findings showed that DE had a high R^2 value and a comparatively low Root Mean Squared Error (RMSE), suggesting that the double exponential smoothing model was more accurate.

Krishnan and Mehta (2024) compared the Artificial Neural Network (ANN) and Exponential Smoothing (ES) models for predicting weather characteristics like wind speed and relative humidity in Kerala's north, middle and south zones Using 36 years (1985-2020) of weather data. MSE, RMSE, MAE, MAPE and R^2 values were used to test the model's correctness. The findings revealed that for wind speed and relative

humidity, the ES model outperformed the ANN in Kerala's central and southern zones, while the ES model outperformed the ANN in the northern zone.

Kumar *et al.* (2024) used Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and Holt-Winters Exponential Smoothing (HWES) models to predict volatility and seasonality vegetation for the three distinct agroclimatic zones in Jharkhand: the central northeastern, eastern and southeastern agroclimatic zones for 20 years of data (2001 to 2021). With optimum values of MAE (16.49), MAPE (0.49), NSE (0.86), RMSE (0.49) and R^2 (0.82), the additive H-W model performed well in the eastern agroclimatic zone, followed by the south-eastern and central north-eastern agroclimatic zones.

12.3 Machine learning model

McCulloch and Pitts (1943) provided the theoretical basis for artificial neural networks by proving that propositional logic may be used to treat neural events and their intricate relationships due to the "all-or-none" nature of nervous activity. The study showed that any logical expression that meets specific requirements can be represented by a similar network. This network successfully simulates the nonlinear functional patterns found in the activity of the nervous system.

Friedman (1991) introduced MARS, an automatic technique for determining product spline basis functions for high-dimensional modeling that employs a forward/backward stepwise procedure. MARS efficiently captures intricate nonlinear interactions by fusing the flexibility of recursive partitioning with the smoothness of spline fitting.

Fausett (1994) discussed a basic introduction to ANN, describing the various applications, techniques and structures that characterize the subject. The book gives a complete introduction to both artificial and biological neural systems. It covers important topics like recognising patterns, processing signals and different ways to train them.

Cortes and Vapnik (1995) worked on A novel learning tool for two-group classification issues were the support-vector network. The following principle was essentially implemented by the machine: input vectors are mapped to a very high-dimensional feature space in a non-linear manner. A linear decision surface is built in this feature space. The support-vector network concept was first used for the limited scenario in which training data separation is error-free.

Suykens and Vandewalle (1999) investigated how to use a linear kernel in Support Vector Regression (SVR) while incorporating linear constraints. They demonstrated that using previous structural knowledge, such as positivity, probability simplex, or monotonicity requirements, preserves the optimization problem's semi-definite quadratic structure.

Breiman (2001) presented the RF algorithm, an ensemble learning technique that uses random feature selection and bootstrap aggregation to merge several decision trees in order to reduce overfitting and increase prediction accuracy. In comparison to conventional single-tree models, the study showed that RF offers superior classification and regression performance, robustness to noise and efficient handling of high-dimensional data.

Bulmer (2003) highlighted Galton's ground-breaking studies on linear regression and trait inheritance, highlighting his transformation from an innovator and explorer to the creator of modern statistical concepts.

Welling (2004) explained the theoretical foundation of Support Vector Regression, which was based on the convex optimization framework and ε-insensitive loss function.

Basak *et al.* (2007) suggested SVR has been used in a various of domains, including convex quadratic programming, loss function selection, approximation of difficult engineering studies, time series and financial (risky and noisy) prediction, etc. An attempt has been made to evaluate the current theory, approaches, new advancements and applications of SVR.

Mekanik *et al.* (2013) used Multiple Regression (MR) and Artificial Neural Networks (ANN) to predict long-term seasonal spring rainfall in Victoria, Australia, by analyzing the concurrent effects. The study revealed that ANN models (with MSEs of 0.008 to 0.018 in the East region, 0.007 to 0.033 in the central region and 0.009 to 0.017 in the West region) that use coupled climatic indices as predictors work much better than MR models. This makes them a better tool for understanding the complex nonlinear interactions that happen in long-term rainfall patterns.

Shao and Hou (2013) developed a two-stage hybrid scheme combining MARS and Support Vector Machines (SVM) to accurately identify change points in multivariate processes. The study utilizes MARS in the first stage to identify and select significant input variables, which are then processed by the SVM in the second stage to

determine the exact time a process shift occurs, effectively reducing computational complexity while maintaining high detection sensitivity.

Deo and Sahin (2015) utilized Artificial Neural Networks (ANN) to predict the monthly Standardized Precipitation and Evapotranspiration Index (SPEI) in eastern Australia using hydrometeorological parameters and climate indices as input variables. The study demonstrates how well the ANN model captures intricate climate patterns to produce accurate drought forecasts in semi-arid areas.

Chen and Guestrin (2016) worked on XGBoost, a scalable end-to-end tree boosting system that makes effective use of weighted quantile sketches and a new sparsity-aware method for approximate tree learning. The study shows that the system can handle billions of cases using out-of-core processing and a cache-aware block structure, far exceeding current techniques on a variety of machine learning benchmarks.

Shiri (2019) assessed the predictive and estimation capabilities of the MARS and RF models for dewpoint temperature using ten years of meteorological data from six Iranian weather stations. Based on the research, prediction-based gene expression programming outperformed both MARS and RF for daily and weekly intervals. However, MARS models were better for estimating values based on air temperature and humidity at the Ahar, Marand and Tabriz stations, with RMSE values of 2.387, 3.003 and 2.421, respectively.

Adnan *et al.* (2020) used temperature and precipitation information from four different input scenarios to assess the predictive performance of MARS and SVM for monthly streamflow prediction in Pakistan's hilly Swat River Basin. The relative advantage of each model depends on catchment characteristics, as confirmed by the fact that SVM reduced the RMSE of the standalone MARS model by 9.12% at Kalam station and MARS reduced RMSE by 2.12% compared to LSSVM at Chakdara station. This finding is crucial for designing MARS-SVR hybrid frameworks, where MARS performs feature reduction before SVR handles the complex nonlinear mapping, as used in the current study for precipitation prediction.

Khairudin *et al.* (2020) investigated using five machine learning models (ANN, SVR, DT, RF and LSTM) with average weekly rainfall data from the Kuala Krai station. In terms of forecasting accuracy, the Random Forest approach outperformed the other models according to RMSE and MAE measures.

Velasco *et al.* (2022) used a dataset spanning more than five years to forecast rainfall in a tropical Philippine city using Support Vector Regression. The study found particular parameter configurations weather values to enhance forecast accuracy by enhancing the Radial Basis Function (RBF) kernel with MSE 3.64.

Dong *et al.* (2023) established an Extreme Gradient Boosting (XGBoost) model-based bias correction technique for Numerical Weather Prediction to improve short-term daily precipitation forecasting in seven distinct climatic zones of China. Using data from 689 weather stations, the XGBoost-based correction greatly decreased forecasting errors and increased the dependability of precipitation estimations.

Das *et al.* (2024) compared seven machine learning models for rainfall prediction, including ANN, RF , MARS and SVR (RMSE 0.0139, 0.071, 0.0722 and 0.0877). The study used different algorithms and data preparation methods to find the best decision support system for rainfall prediction. It relied on meteorological data, including temperature, relative humidity, wind direction and speed.

Kashefi *et al.* (2026) utilized Decision Tree, Random Forest and XGBoost algorithms to model monthly and maximum 24-hour precipitation for Lamerd, Iran, using a 28-year dataset of meteorological variables. The study found that while all models performed satisfactorily, the XGBoost model demonstrated superior accuracy (RMSE training 9.93 and testing 11.24) in predicting precipitation patterns compared to the other two algorithms RF (RMSE training 11.07 and Testing 13.31) and XGBoost (RMSE training 10.75 and Testing 12.80).

12.4 Stochastic model

Javadinejad *et al.* (2023) used first-order and second-order Markov chain stochastic models to investigate the monthly drought in the southern region of Iran between 1980 and 2020. The majority of the stations had normal, moderate and severe drought conditions, based on the results. The transition probability matrix also indicated that all stations had a high chance of moving from one state to another and from a wet to a dry state but a low chance of moving from a dry to a wet state.

Chowdhury *et al.* (2025) used ML models with the first, second and third Markov chain models to estimate rainfall in four Indian cities, viz., Bhubaneshwar, Pune, Bangalore and Hyderabad for the period of 1969 and 1999. Confusion matrices and rainfall statistics were combined to assess how well the different models simulated rainfall events. The outcome showed that the machine learning model performed better than the Markov chain model.

13. Materials and Mehods

13.1 Source of data

The weather data related to study area will be analysed for different stations under Junagadh Agricultural Universities of Gujarat.

13.2 Description of Data

The basic statistical parameters of the monthly, seasonal and annual precipitation series will be computed, including the arithmetic mean, standard deviation, coefficient of variation, skewness, kurtosis, minimum, maximum and median. These measures provide an initial characterization of the central tendency, dispersion and distributional shape of the precipitation data across different temporal aggregation levels.

Monthly and seasonal mean rainfall distributions will be examined to understand the intra-annual pattern of precipitation concentration, with particular focus on the south-west monsoon season (June to September). The contribution of monsoon rainfall to total annual rainfall will be computed year-wise and expressed as a percentage for trend assessment.

Frequency analysis of weekly rainfall amounts will be performed by fitting appropriate probability distribution functions to the observed weekly rainfall series. The distribution fitting and goodness-of-fit assessment will be used the Chi-square and Kolmogorov-Smirnov tests.

13.3 RESEARCH METHODOLOGY

This section will be divided into three sections: (1) determining the trend of weather parameters; (2) short-term precipitation forecasting using univariate time series analysis, such as exponential smoothing models, as well as multivariate time series, such as machine learning models; and (3) long-term precipitation forecasting using stochastic Markov chain models. In trend analysis, the Mann-Kendall and Sen's Slope nonparametric tests will be employed to determine the trend for feature variables. In univariate time series analysis, double and triple exponential smoothing techniques will be used to predict rainfall. For multivariate analysis, different machine learning techniques *viz.,* MLR, MARS, ANN, SVR, RF, XGBoost and Hybrid model (MARS-ANN and MARS-SVR) will be used taking rainfall as dependent variable and remaining feature variables as an input feature. The comparison of all these machine learning algorithms will be done based on evaluation metrics. The long term

precipitation forecasting will be done by using first order Markov chain model approach.

13.3.1 Initial Assessment

For every meteorological parameter, descriptive statistics will be first performed, including the number of observations, missing values, minimum, maximum, quartiles, average, standard deviation and coefficient of variation. The KNN (K Nearest Neighbor) imputation method will be utilized to fill in the missing variables.

13.3.2 KNN Imputation

A popular technique for imputing missing data during the data preprocessing step is the KNN imputation method. This phase has a major effect on the model's performance Altamimi *et al.*, (2024).

KNN imputation can be performed using a variety of distance measures, including the Euclidean, Minkowski, Manhattan, Cosine, Jaccard and Hamming distances. Nonetheless, it is said that the Euclidean distance is the most effective and fruitful.

The Euclidean distance is used to find out KNN imputation as follows Emmanuel *et al.*, (2021).

$$Dist_{xy} = \sqrt{\sum_{j=1}^{k}(X_{ik} - X_{jk})^2} \quad \textbf{...Eq. 13.1}$$

Where, $Dist_{xy}$ is euclidian distance, k is data attributes, j is 1, 2, 3 ... k , X_{ik} = value for ith attribute containing missing data and X_{jk} is value of jth attribute containing complete data

Weight Mean Estimation will be computed after selecting value of the k points with the lowest distance.

$$X_k = \frac{\sum_{j=1}^{j} w_j v_j}{\sum_{j=1}^{j} w_j} \quad \textbf{...Eq. 13.2}$$

Where, X_k is mean estimation, J is the number of parameters used with j = 1,2,3…K, v_j is complete values on attributes containing missing data, w_j is nearest neighbors observed.

$$w_j = \frac{1}{dis_{(x,y)^2}} \quad \textbf{...Eq. 13.3}$$

The R or Python programming language will be used for missing imputation.

14. Statistical Analysis

The following statistical tools will be used to achieve objectives of the study.

14.1 Weather Trend Analysis

Mann Kendall and Sen's slope approach will be used for the trend analysis of the month-wise weather parameters.

14.1.1 Mann-Kendall test

In hydro-meteorological processes, the non-parametric Mann-Kendall (MK) test (Mann 1945; Kendall 1975) is frequently employed to identify important trends (Halder *et al.*, 2023).

The MK statistics will be calculated using:

$$S = \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} sig(x_i - x_j) \quad \textbf{...Eq. 14.1}$$

Where, n is no. of data point, x_i and x_j is data value in time series and $sig(x_i - x_j)$ is the function of sign was given below.

$$sig(x_i - x_j) = \begin{bmatrix} -1, & if\ (x_i - x_j) < 0 \\ 0, & if\ (x_i - x_j) = 0 \\ +1, & if\ (x_i - x_j) > 0 \end{bmatrix} \quad \textbf{...Eq. 14.2}$$

The variance of S was calculated by using

$$var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{k} p_i(p_i-1)(2p_i+5)}{18} \quad \textbf{...Eq. 14.3}$$

Where, k is no. of tied group, p_i is no. of data value in kth group.

When n > 10, the standard normal test statistic Z_s will be computed using

$$Z_s = \begin{bmatrix} \frac{S-1}{\sqrt{var(S)}}, & if\ S > 0 \\ 0, & if\ S = 0 \\ \frac{S+1}{\sqrt{var(S)}}, & if\ S < 0 \end{bmatrix} \quad \textbf{...Eq. 14.4}$$

A declining trend is shown by a negative Z_s value, whereas a rising trend is indicated by a positive Z_s value.

14.1.2 Sen's Slope (SS) estimator

The magnitude of the trend in a time series will be estimated using this non-parametric method (Sen, 1968). The trend's slope, denoted by Q_i, was calculated as follows:

$$Q_i = \frac{t_i - t_k}{j-k} \quad \textbf{...Eq. 14.5}$$

Where, t_i and t_k = data values at j and k time

Sen's slope estimator, often known as the median slope, will be computed by listing the N values of Q_i from least to greatest.

$$Q_{med} = \begin{bmatrix} Q_{\frac{N+1}{2}}, & if\ N\ is\ odd \\ \frac{Q_{\frac{N}{2}} + Q_{\frac{N+1}{2}}}{2}, & if\ N\ is\ even \end{bmatrix} \quad \textbf{...Eq. 14.6}$$

A positive value of Q_{med} denotes an upward trend in the time series, whereas a negative value of Q_{med} suggests a downward trend.

Sen's slope approach and the Mann Kendall non-parametric test were found to be the most helpful in the majority of weather trend research investigations. Sen's slope method can be used to determine the slope of weather parameters, whereas the Mann Kendall statistics test can be used to determine the trend of weather parameters. Therefore, the Mann Kendall statistical test will be utilized in this investigation.

14.2 Time Series Analysis

Univariate time series forecasting is a method for predicting future values of a single variable using its previous data. Rainfall will be predicted utilising univariate time series forecasting models, such as the double exponential model and the triple exponential smoothing models. Monthly rainfall data will be used in this univariate time series.

14.2.1 Double exponential smoothing

When there is a trend in the data, double exponential smoothing will be applied. This approach updates the trend and seasonal components for every period (Holt, 1957).

For level:

$$L = \alpha(x_t) + (1 - \alpha)(L_{t-1} + b_{t-1}) \quad \textbf{...Eq. 14.7}$$

For trend:

$$b_t = \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \quad \textbf{...Eq. 14.8}$$

$$F_t = L_t + b_t \quad \textbf{...Eq. 14.9}$$

Where, b_t is trend component at time t, b_{t-1} = trend at time t-1, x_t is data for time period, L_t is level component at time t, L_{t-1} is level at time t-1, F_t is forecast value, α is smoothing parameter for level and β is smoothing parameter for trend.

14.2.2 Holt Winter method

A technique for identifying trends in time series forecasting was created by Holt in 1957 and expanded by Winters in 1960 to take seasonal fluctuations into account.

The forecasting model uses three smoothing equations: level (L_t), trend (b_t) and seasonal component (S_t). These equations are changed using (α, β and γ) smoothing parameters, which give recent data more weight. The level is represented by the smoothing parameter α, the trend by β and the seasonal component by γ. The frequency

of the seasonal cycle, or the number of seasons in a year, is represented by the parameter m.

There are two kind of models in the Holt Winter method: multiplicative and additive. The additive method is appropriate when seasonal variations are consistent over time. The multiplicative method is employed when seasonal changes are proportionate to the level of the series. If there are some zero values in the rainfall dataset, then multiplicative model will not be applied.

Additive Holt Winters Method (AHW) (Lallawmzuali and Muniyan, 2025)

The Additive Holt Winters (AHW) method will be used when time series data has a stable seasonal part over time. Four equations are used by the AHW technique to calculate the series' historical average of the series (Level), predict future changes (Trend) and account for the seasonal pattern (Seasonality). The forecast is calculated as a weighted total of these earlier terms in the final equation, which is as follows:

$$L_t = \alpha\,(Y_t - S_{t-m}) + (1 - \alpha)(L_{t-1} + B_{t-1}) \quad \textbf{...Eq. 14.10}$$

$$B_t = \beta\,(L_t - L_{t-1}) + (1 - \beta)B_{t-1} \quad \textbf{...Eq. 14.11}$$

$$S_t = \gamma(Y_t - L_t) + (1 - \gamma)S_{t-m} \quad \textbf{...Eq. 14.12}$$

$$F_{t+h} = (L_t + B_t h) + S_{t-m+h} \quad \textbf{...Eq. 14.13}$$

$$Ft+h = (Lt + Bth) + St-m+h \quad \textbf{...Eq. 14.14}$$

where, m is seasonality length, L_t is level components, B_t is trend component, S_t is seasonal component, Y_t is data for time period and F_{t+h} is forecast for the period ahead (h).

α, β and γ are the smoothing parameters for the level, trend and seasonality, respectively and range is 0 to 1.

For the current study, the seasonal length, m=12 as working with monthly data.

14.3 Machine Learning Model

The current study will use machine learning algorithms for short-term precipitation forecasting and compared these methods using assessment measures.

14.3.1 Data preprocessing

To achieve the best model performance, machine learning techniques require data preprocessing (Kotsiantis *et al.*, 2006). It guarantees that the dataset is pure, reliable and appropriate for training models. The following data preprocessing techniques will be applied to get the dataset ready for machine learning model-based precipitation prediction.

Standardization

Standardization is required for this data in order to address the varied units used for weather parameters. All data is transformed into mean zero and unit variance using this standardization scalar. The data will be transformed using the Standard Scalar technique, which does the following:

$$z = \frac{x_i - \mu}{\sigma} \quad \textbf{...Eq. 14.15}$$

where, x_i is i[th] observation, μ is Mean value, σ is Standard deviation.

This process gives no single feature variable disproportionately to the model.

Data split

For effective model building, the dataset will be split into 70:20:10 or 80:10:10 training, testing and validation datasets.

14.3.2 Multiple Linear Regression model

The multiple linear regression analysis will be employed find out the relative contribution of the independent variables to the dependent variable.

In the stepwise method, Regression of y and $x_i \ldots x_k$ taken separately will the first step in the regression analysis. The first variable chosen is the one with the highest accountability in the sum of squares of y. The bivariate regression that included x_i will be calculated. After fitting the x_i variable, the variable that provides the greatest additional accountability in the sum of squares in y was chosen. Every trivariate regression involving both x_1 and x_2 will be calculated. The final variable, whose new contribution was the least of all the factors, was the subject of further analysis.

The prediction equation used as:

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \quad \textbf{...Eq. 14.16}$$

Where, y is Dependent variable, a is Intercept, $b_1 \ldots b_k$ is Partial regression co-efficient of respective independent variables, $x_1 \ldots x_k$ is Independent variables.

After the regression equation, the 'F' values for partial regression co-efficients will be tested for their significance.

14.3.3 Multivariate Adaptive Regression Splines model

Friedman (1991) developed a nonparametric method called multivariate adaptive regression spline. The primary objective of this technique is to forecast the values of a continuous dependent variable, $y(n \times 1)$, based on a collection of independent explanatory variables, $X(n \times p)$. The MARS model can be stand as:

$$y = f(X) + e \quad \textbf{...Eq. 14.17}$$

where, e is a dimension-based error vector. $(n \times p)$. $f(x)$ is basically sum of intercept and basic functions generated by MARS algorithm.

The foundation of MARS is local regression modeling. It approximates complex nonlinear relations using spline functions. The entire data region is divided into many subdomains by MARS. A regression line is used to fit a response variable at each subdomain. A polynomial function makes up every part of a spline of degree q. MARS uses two-sided truncated power functions as the basis for splines. This function can be described by the following equations (Sekulic and Kowalski, 1992):

$$-(x-t)_+^q = \begin{cases}(t-x)^q, if\ x < t, \\ 0,\ otherwise\end{cases} \quad \textbf{...Eq. 14.18}$$

$$+(x-t)_+^q = \begin{cases}(t-x)^q, if\ x \geq t, \\ 0,\ otherwise\end{cases} \quad \textbf{...Eq. 14.19}$$

Where, $q(\geq 0)$ is the power of splines to determine the degree of smoothness of the resultant function estimate

The model MARS for a dependent (outcome) variable y and M terms, can be expressed in the following equation (Friedman and Roosen, 1995):

$$\hat{\hat{y}} = \hat{f}_M(\vec{X}) = c_0 + \sum_{m=1}^{M} c_m B_m(\vec{X}) \quad \textbf{...Eq. 14.20}$$

Where, $\hat{\hat{y}}$ = the dependent variable predicted by the MARS model c_0 is a constant, $B_m(\vec{X})$ is the m^{th} basis function and c_m is the coefficient of the m^{th} basis function.

There may be just one spline basis function that makes up the basis function. We need to find the best values for both the variables that will be added to the model and the knot positions for each variable. For a data set $\vec{X}$ containing n objects and p explanatory variables, there are $N = n \times p$ pairs of spline basis functions, given by Eq. (14.18) and (14.19), with knot locations x_{ij} = 1, 2, . . ., n; j = 1, 2, . . ., p). In the present study, x=148 breeding lines will collection i.e. p=7 and n=148 and x = 465 weather data point will collection i.e. p = 9 and n = 465.

The final model will be built using a two-step process. First, the subsequent pairs of the model's basis functions will be chosen using a two-at-a-time forward stepwise process (Friedman and Roosen, 1995). This forward stepwise selection process leads to a very complex and over-fitted model. The fitted model has poor predictive abilities for new objects.

Second a backward stepwise procedure will be used to improve the prediction. This procedure removed the redundant basis functions one at a time. Moreover, to determine which basis functions should be included in the model.

MARS utilizes Generalized Cross-Validation (GCV) (Sekulic and Kowalski, 1992). The GCV is the mean squared residual error divided by a penalty which is dependent on model complexity. The GVC criterion is defined as follows

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{N}(y_i - f_M(\overrightarrow{X_i}))^2}{(1-\frac{C(M)}{n})^2} \quad \text{...Eq. 14.21}$$

where *C(M)* is a complexity penalty that increases with the number of basis functions in the model and which is defined as:

$$C(M) = (M+1) + dM \quad \text{...Eq. 14.22}$$

where *M* is the number of basis functions and the parameter *d* is a penalty for each basis function included into the model. *d* is a smoothing parameter.

14.3.4 Artificial Neural Network model

An efficient model for prediction nonlinear time series data is Artificial Neural Networks (ANN) (Zhang *et al.*, 1998; Jha *et al.*, 2009). Because they mirror the behaviour of the human brain's central nervous system, these models are referred to as "Neural Network" models. They belong to a class of data-driven, nonparametric and generalized nonlinear statistical techniques. Thus, they are ideally suited for the modelling of agricultural data which are known to be complex and often non-linear. Fig. 14.1 displays a block schematic of a neuron model based on the architecture of an artificial neural network.

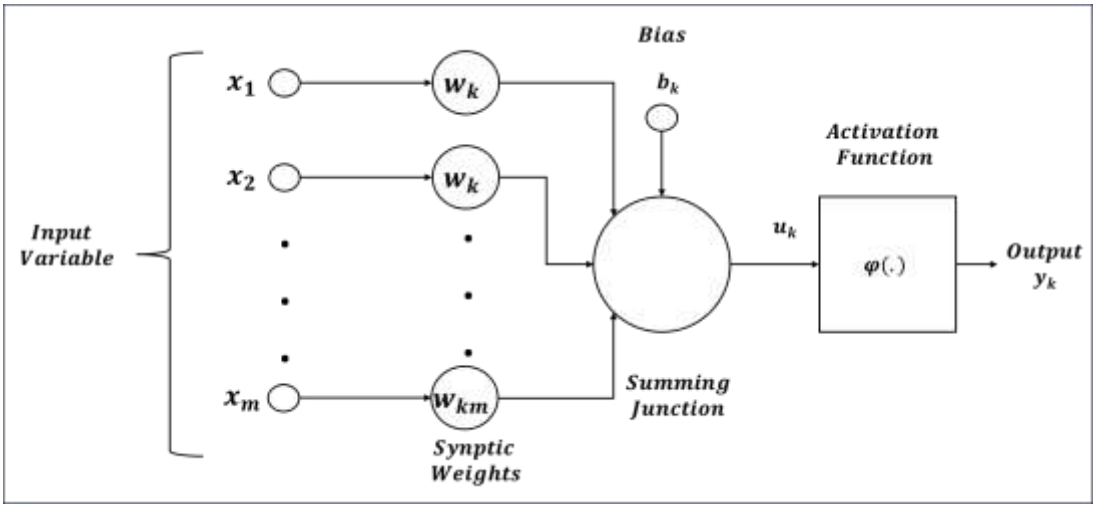

Figure 14.1 Block diagram of model of a neuron

An input layer that utilizes in external data, one or more hidden layers that give the model nonlinearity and an output layer that produces the desired value make up the

design. Given a sufficient number of hidden nodes and sufficient training data points, a neural network with one hidden layer may approximate any nonlinear function.

The foremost thing in model fitting is dividing the dataset into training and testing. Most of the researchers follow particular procedures for splitting dataset. That is use 80% or 90% for training and 20% or 10% for testing. In this study, last 20 percent data will be used as testing.

Architecture of ANN

The ANN model is made up of multiple interconnected units called neurons. One input layer, one output layer and one or more hidden layers make up a standard ANN model. Each layer has nodes, which include input nodes, output nodes and hidden nodes. Each layer's output is the weighted sum of its inputs, which is a typical feature.

According to Haykin (1999), the following equations can be used to mathematically define a neuron k:

$$u_i = \sum_{j=1}^{m} w_{kj} x_j \quad \textbf{...Eq. 14.23}$$

$$y = \varphi(u_k + b_k) \quad \textbf{...Eq. 14.24}$$

Here bias (b_k), has the effect of increasing or lowering the net input of the activation function. $x_1, x_2, \ldots, x_m$ are the inputs; $w_{k1}, w_{k2}, \ldots, w_{km}$ are the weights of the neuron k; u_k is the linear combiner output due to input variables; $\varphi(.)$ is the activation function; y_k is the output of the neuron the weight attached to the connection from j^{th} hidden node to the output node.

Sigmoid or Logistic activation function: $(x) = \frac{1}{1+e^{-x}}$; $rang\ (0,1)$ **...Eq. 14.25**

Activation function is a differentiable function that is used for smoothing the result of the cross product of the covariate or neurons and the weights. The logistic function will be employed as an activation function in the current study to modify the weights in the feedforward networks with several layers.

14.3.5 Support Vector Regression model

Vapnik *et al.* (1997) invented the Support Vector Machine. In machine learning, support vector machines, or SVMs, are supervised learning models with corresponding learning algorithms that analyze data for regression analysis and classification. The two types of SVM are Support Vector Regression (SVR) and Support Vector Classification (SVC).

SVM is a high-dimensional feature space learning system. It creates prediction functions based on a selection of support vectors. SVM is a revolutionary compression

technique that only few support vectors to generalize intricate grey-level structures. An SVM for regression was published in 1997 by Vapnik, Steven Golowich and Alex Smola (Vapnik *et al.*, 1997). Support vector regression (SVR) is the name given to this method. Because training points outside the margin are irrelevant to the cost function that was used to build the model. Only a portion of the training data is used to build the support vector classification model.

SVM is primarily utilized in situations when the data are not separable linearly. To classify such a linearly separable data set, the data are transformed into a higher-dimensional space using transformation functions (kernel functions), where the data may be separable linearly. This not only classifies the data but also optimizes the decision border by increasing the margin between data clusters. Similarly, because the cost function for generating the model eliminates any training data close (within a threshold) to the model prediction, SVR only relies on a subset of the training data.

The accuracy of SVR models depends on how well the loss function, error penalty factor C and SVR meta-parameters are configured. Additionally, the final models will be significantly impacted by the choice of the kernel function. In this study, radial basis kernel function (RBF) will be used, $k(x, x') = exp(-(\frac{|x-x'|^2}{2\sigma^2}))$. SVR model will be created with C as 1, ε as 0.1 and 0.1 for the RBF kernel parameter.

Fig. 14.2 depicts an SVR functioning diagram. In SVR, we aim to consider the points within the decision boundary line. Consider the two red lines as the decision boundary and the green line as the hyperplane will be considered. The hyperplane with the most points is the best fit line.

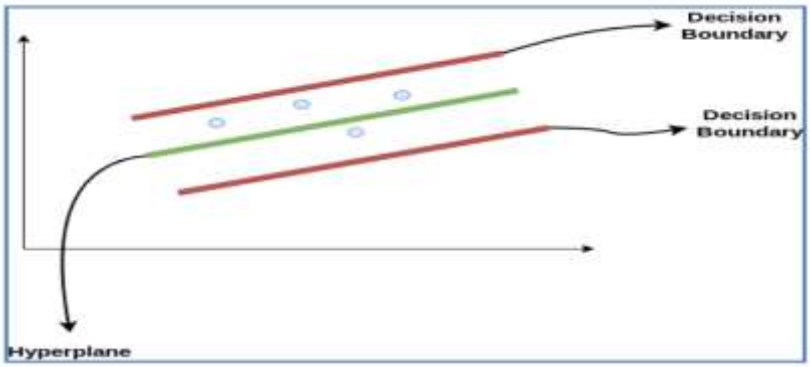

Figure 14.2 Diagram of SVR algorithm

14.3.6 Random Forest model

Random Forest is based on the Decision Trees (DT) principle, but it does not select all data points and variables in each tree (Breiman, 2001). It creates trees by

randomly choosing data points and variables and combining the results. Additionally, it eliminates any bias that may have been introduced into the system.

Furthermore, it considerably improves predictive power. The RF technique is ensemble-based and enhances model accuracy. Due to the randomness, the RF algorithm examines all variables at each node to determine how well they split. Generally, RF outperforms all other machine learning algorithms in processing speed and resistance to overfitting noise. According to (Montantes, 2020), RF is an ensemble of decision trees. The final/leaf node is either the majority class for classification issues or the average for regression problems. By merging several trees into one ensemble model, an RF helps lower the high variance of a flexible model like a decision tree.

Decision Tree

A decision tree regression algorithm creates a graph flowchart to depict a decision's various outcomes. To begin building the tree, the algorithm determines which variable best divides the data into two groups. It then repeats the preceding process with the remaining variables. As a result, a tree graph is created, with each break representing a decision. The algorithm chooses splits so that the maximum number of observations is correctly categorized. The primary advantage of a decision tree is that it is intuitive and easily comprehended by anyone with no prior knowledge of the subject (Magee, (1964)).

14.3.7 Extreme Gradient Boosting model

XGBoost, or eXtreme Gradient Boosting, is an advanced implementation of the gradient boosting framework that has become widely recognized for its speed, efficiency and accuracy. It works by sequentially combining multiple weak learners, typically decision trees, in a way that each new tree focuses on reducing the errors made by the previous ensemble (Friedman, 2001). This iterative error-correcting mechanism makes XGBoost highly effective for capturing complex, non-linear relationships within data. Unlike conventional boosting methods, XGBoost integrates several enhancements such as regularization to control overfitting, parallelized tree construction, handling of missing values and pruning strategies, which together contribute to improved model robustness and computational efficiency. Owing to these advantages, XGBoost has found extensive use in tasks such as regression, classification and ranking and is considered one of the most reliable algorithms for large-scale structured datasets.

Incorporation of both first-and second-order derivatives of the loss function during optimization is the distinctive feature of XGBoost, which allows more precise

approximation of gradients and faster convergence (Chen and Guestrin, 2016). In addition, the algorithm introduces regularization terms into the objective function, ensuring that models remain generalizable rather than overly complex. Hyperparameters such as the learning rate (η), tree depth, number of estimators, subsample ratios and column sampling ratios give users fine-grained control over model flexibility and bias-variance trade-offs. The combination of these design choices makes XGBoost not only accurate but also highly adaptable to diverse prediction problems.

The objective function minimized by XGBoost is expressed as:

$$L(\varphi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad \textbf{...Eq. 14.26}$$

Where $l(y_i, \hat{y}_i)$ measures the difference between observed and predicted values and $\Omega(f_k)$ represents the regularization term for each tree. The regularization is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \quad \textbf{...Eq. 14.27}$$

With T as the number of leaves, as the w_j weight of leaf j, γ as the penalty for additional leaves and λ as the L2 regularization parameter. To optimize this objective, XGBoost applies a second-order Taylor expansion of the loss function using the gradient and Hessian:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}} \quad \textbf{...Eq. 14.28}$$

Where g_i and h_i denote the first- and second-order derivatives of the loss at iteration *t - 1*. The optimal weight for a leaf node is then given by:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad \textbf{...Eq. 14.29}$$

and the gain from a candidate split is calculated as:

$$gain = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \textbf{...Eq. 14.30}$$

Where I_L and I_R are the data instances assigned to the left and right child nodes, respectively. A split is accepted only when the gain is positive, ensuring improved predictive accuracy. In this study, hyperparameters were tuned using cross-validation and early stopping was applied to prevent overfitting, with model performance evaluated through appropriate statistical measures depending on the prediction task.

14.3.8 Hybrid model (MARS-SVR and MARS-ANN)

In the given study, MARS based hybrid models will be developed by combining it with ANN and SVM. The basic idea is the variable selection ability of MARS algorithm and prediction ability of ANN/SVM. Firstly, important variables among the independent variables will be extracted by MARS algorithm. Then these selected variables will be taken as input variables to predict dependent variable.

MARS model: $\hat{\tilde{y}} = c_0 + \sum_{m=1}^{M} c_m B_m(\vec{X})$ **...Eq. 14.31**

Where $\hat{\tilde{y}}$ is the dependent variable predicted by the MARS model, c_0 is a constant, $B_m(\vec{X})$ is the mth basis function and c_m is the coefficient of the m^{th} basis function.

Modified ANN model: $\tilde{u}_i = \sum_{j=1}^{m} w_{kj}\tilde{x}_j$ **...Eq. 14.32**

$\tilde{y}_k = \varphi(\tilde{u}_k + b_k)$ **...Eq. 14.33**

Here $\tilde{x}_j$ are the selected important variables through MARS. $\tilde{u}_k$ is the linear combiner output due to input variables; $\varphi(.)$ is the activation function; $\tilde{y}_k$ is the output of the neuron the weight attached to the connection from j^{th} hidden node to the output node.

Modified SVM model: $\tilde{y} = \phi(\tilde{x}) = W\varphi(\tilde{x}) + b$ **...Eq. 14.34**

Here $\tilde{x}_j$ are the selected important variables through MARS. $\tilde{y}_k$ is the output of modified SVM model. w defines weight vector, φ denotes mapping function and b is a bias term. Schematic diagram was also presented in Fig. 14.3.

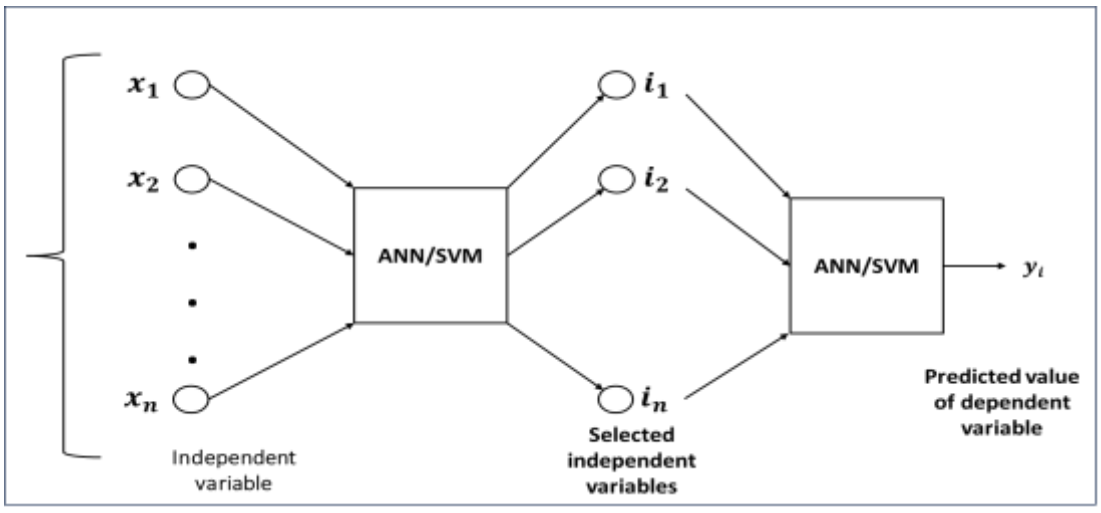

Figure 14.3 MARS based ANN/SVM hybrid framework

14.3.9 Model Evaluation and Comparison

All eleven models (MLR, MARS, ANN, SVR, RF, XGBoost, MARS-ANN and MARS-SVR) will be evaluated on the independent test dataset using RMSE, MAE and

R^2. The model with lowest RMSE, MAE and highest R^2 will be identified as the best performing model for each of the three prediction tasks.

The formula for RMSE is given by,

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad \textbf{...Eq. 14.35}$$

MAE is given by,

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \quad \textbf{...Eq. 14.36}$$

Coefficient of Determination (R^2) is given by,

$$R^2 = \frac{RSS}{TSS} \quad \textbf{…Eq. 14.37}$$

Where, y_t is the actual value for time t, $\hat{y}_t$ is the predicted value for time t, n is the number of observations used in predictions, RSS is Regression Sum of Squares, ESS is Error Sum of Squares and TSS is Total Sum of Squares.

14.4 Stochastic Markov Chain Model

Markov chains is stochastic models that show a series of events in which each event's probability is solely dependent on the state achieved in the previous event. Markov chain models are used in rainfall to generate realistic rainfall patterns for hydrological and meteorological investigations by capturing the temporal dependency structure of rainfall occurrences.

Every succeeding value in a first-order Markov chain depends just on the state that came before it. The next state in a Markov chain of second or higher order depends on two or more previous states (Shamshad *et al.*, 2005).

Let $X(t)$ be a stochastic process, possessing a discrete state space $S = \{1, 2, \dots, K\}$. In general, for a given sequence of time points $t_1 < t_2 < \dots < t_{n-1} < t_n$, the conditional probabilities are as following.

$$Pr\{X(t_n) = i_n \mid X(t_1 = i_1, \dots, X(t_{n-1}) = i_{n-1}\} = Pr\{X(t_n) = i_n | X(t_{n-1}) = i_{n-1}\} \quad \textbf{…Eq. 14.38}$$

The conditional probabilities $Pr\{X(t) = j|X(s) = i\} = P_{ij}(s,t)$ are called transition probabilities of order $r = t - s$ from state i to state j for all indices $0 \leq s < t$, with $1 \leq i$ and $j \leq k$. They are denoted as the transition matrix P. For k states, the first order transition matrix P has a size of $k \times k$ and takes the form:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \quad \textbf{…Eq. 14.39}$$

The relative probabilities of the k states can be used to determine the state probable at time t. If the number of transitions from state i to state j in the data sequence is nij, then the maximum likelihood estimates of the transition probabilities is

$$p_{ij} = \frac{n_{ij}}{\sum_j n_{ij}} \quad \textbf{...Eq. 14.40}$$

The transition probabilities of any state vary between 0 and 1. The summation of transition probabilities in a row equals one. Mathematically, it can be expressed as:

$$\sum_{j=1} p_{ij} = 1 \quad \textbf{...Eq. 14.41}$$

For the current study, eight different classifications of rainfall events are presented in Table 14.1 (Sonar, 2014).

Table 14.1: Classification of rainfall events

Sr. No	Category	Rainfall (mm/day)
1	No rain	0.0
2	Very light rain	0.1 to 2.4
3	Light rain	2.5 to 7.5
4	Moderate rain	7.6 to 35.5
5	Rather heavy rain	35.6 to 64.4
6	Heavy rain	64.5 to 124.4
7	Very heavy rain	124.5 to 244.4
8	Extremely Heavy Rain	> 244.5

Month by month analysis of rainfall data of monsoon season for long term will be carried out using first order Markov chain model. The methodology will be carried out with forming transition matrix and computing probabilities for each month.

15. References

Adnan, R. M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O. and Li, B. (2020). Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. *Journal of Hydrology*, **586**: 124371.

Altamimi, A., Alarfaj, A. A., Umer, M., Alabdulqader, E. A., Alsubai, S., Kim, T. H. and Ashraf, I. (2024). An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques. *BMC Medical Research Methodology*, **24(1)**: 221.

Anonymous (2025). Ministry of Statistics and Programme Implementation. Press note on provisional estimates of annual gross domestic product for 2024–25 and quarterly estimates of gross domestic product for the fourth quarter (January–March) of 2024–25. Government of India. p. 3.

Basak, D., Pal, S. and Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, **11(10)**: 203-224.

Breiman, L. (2001). Random forests. *Machine learning*, **45(1)**: 5-32.

Bulmer, M. G. (2003). *Francis Galton: Pioneer of Heredity and Biometry*. JHU Press, Baltimore. p. 306.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In:* Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, USA, August 13-17, 2016. pp. 785-794.

Chowdhury, D. P., Roy, D. and Saha, U. (2025). Study of Rainfall Occurrence Process by Markov Chain Models and Decision Tree-based Ensemble and Boosting Techniques. *Water Resources Management*, **39(6)**: 2857-2877.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, **20(3)**: 273-297.

Das, P. K., Sahu, R. L. and Swain, P. C. (2024). Comparative analysis of machine learning models for rainfall prediction. *Journal of Atmospheric and Solar-Terrestrial Physics*, **264**: 106340.

Das, P., Jha, G. K., Lama, A. and Parsad, R. (2023). Crop yield prediction using hybrid machine learning approach: a case study of lentil (Lens culinaris Medik.). *Agriculture*, **13(3)**: 596.

Deo, R. C. and Şahin, M. (2015). Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia. *Atmospheric research*, **161**: 65-81.

Dong, J., Zeng, W., Wu, L., Huang, J., Gaiser, T. and Srivastava, A. K. (2023). Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China. *Engineering Applications of Artificial Intelligence*, **117**: 105579.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big data*, **8(1)**: 140.

Fausett, L. (1994). *Fundamentals of Neural Networks.* Prentice Hall International, Upper Saddle River, New Jersey, USA. pp. 3-6.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, **19(1)**: 1-67.

Friedman, J. H. and Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Statistical methods in medical research*, **4(3)**: 197-217.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *In*: Annals of statistics, Institute of Mathematical Statistics, Ohio, USA, pp. 1189-1232.

Goel, S. and Singh, R. K. (2025). Markov Chain Analysis and Incomplete Gamma Probability Function for Rainfall Distribution in the Plain and Hill Zones of Uttarakhand, India. *Journal of Geography, Environment and Earth Science International*, **29(3)**: 82-95.

Halder, P., Dey, R. K. and Mandal, S. (2023). Long-period trend analysis of annual and seasonal rainfall in West Bengal, India (1901–2020) Long-period trend analysis of annual and seasonal rainfall in West Bengal, India (1901–2020). *Theoretical and Applied Climatology*, **154(1)**: 685-703.

Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Macmillan College Publishing, New York, USA. Prentice hall PTR, New Jersey. pp. 696.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, **20(1)**: 5-10.

Javadinejad, S., Dara, R., Krause, S., Riyahi, M., Mohammadi, A. and Nasseri, M. (2023). Analysis and forecasting drought characterisation with using Markov chain model. *Safety in Extreme Environments*, **5(3)**: 149-159.

Jha, G. K., Thulasiraman, P. and Thulasiram, R. K. (2009). PSO based neural network for time series forecasting. *In:* 2009 International Joint Conference on Neural Networks, San Diego, CA, USA, June 14-19, 2009. IEEE Computational Intelligence Magazine, Robert Kozma (ed.), University of Memphis, Los Alamitos, USA. pp. 1422-1427.

Joshi, G. S., Chandole, V., Patel, P., Rana, S. C. and Srivastava, V. K. (2023). Spatio-temporal trend assessment of extreme temperatures and association with land use dynamics and water resource management practices in Bhogavo River watersheds, Gujarat, India. *Sustainable Water Resources Management*, **9(3)**: 63.

Kashefi, M., Karami, H., Niksefat, M. and Ghazvinian, H. (2026). Prediction of monthly precipitation and maximum 24 h precipitation using Random Forest, Decision Tree and XGBoost models. *Modeling Earth Systems and Environment*, **12(1)**: 70.

Kendall, M. G. (1975). Rank correlation measures, *Charles Griffin*, London. pp 202.

Khairudin, N. B. M., Mustapha, N. B., Aris, T. N. B. M. and Zolkepli, M. B. (2020). Comparison of machine learning models for rainfall forecasting. *In:* 2020 international conference on computer science and its application in agriculture, Bogor, Indonesia, September 16-18, 2020. IEEE Indonesia Section, Indonesia. pp. 1-5.

Kotsiantis, S. B., Kanellopoulos, D. and Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, **1(2)**: 111-117.

Krishnan, K. B. and Mehta, V. (2024). Comparison study on modelling and prediction of weather parameters combining exponential smoothing and artificial neural network models in different zones of Kerala. *Environment and Ecology,* **42 (3)**: 1094—1103.

Kumar, N., Panchal, C. C., Chandrawanshi, S. K. and Thanki, J. D. (2017). Analysis of rainfall by using Mann-Kendall trend, Sen's slope and variability at five districts of south Gujarat, India. *Mausam*, **68(2)**: 205-222.

Kumar, V., Bharti, B., Singh, H. P., Singh, A. and Topno, A. R. (2024). Prediction of volatility and seasonality vegetation by using the GARCH and Holt-Winters models. *Environmental Monitoring and Assessment*, **196(3)**: 288.

Lallawmzuali, M. and Muniyan, S. (2025). Rainfall prediction in Mizoram using the modified Holt Winters method. *Theoretical and Applied Climatology*, **156(2)**: 125.

Li, X., Li, C., Guo, F., Meng, X., Liu, Y. and Ren, F. (2024). Coefficient of variation method combined with XGboost ensemble model for wheat growth monitoring. *Frontiers in plant science*, **14**: 1267108.

Lodhavi, M. (2022). Temporal Patterns and Variability of Rainfall Over the Saurashtra-Kutch Region of Gujarat (1901–2017). *In* : International Geographical Union Thematic Conference, Singapore, November 24-25, 2022. Springer Nature Singapore, Singapore, pp. 302-315.

Magee, J. F. (1964). Decision trees for decision making. *In: Harvard Business Review.* Harvard Business Review, Brighton, MA, USA. pp. 35-48.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the econometric society*, **1**: 245-259.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5(4)**: 115-133.

Mehta, D. and Yadav, S. M. (2021). Analysis of long-term rainfall trends in Rajasthan, India. *In*: Climate Change Impacts on Water Resources: Hydraulics, Water Resources and Coastal Engineering, Cham, December 23-25, 2021. Springer International Publishing, Cham, Switzerland, pp. 293-306).

Mehta, D., Waikhom, S., Yadav, V., Lukhi, Z., Eslamian, S. and Furze, J. N. (2022). Trend analysis of rainfall: A case study of Surat City in Gujarat, Western India. *In*: *Earth Systems Protection and Sustainability.* Cham: Springer International Publishing, Cham, Switzerland, *Volume 2,* pp. 191-202.

Mekanik, F., Imteaz, M. A., Gato-Trinidad, S. and Elmahdi, A. (2013). Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology*, **503**: 11-21.

Mondal, A., Kundu, S. and Mukhopadhyay, A. (2012). Rainfall trend analysis by Mann-Kendall test: A case study of north-eastern part of Cuttack district, Orissa. *International Journal of Geology, Earth and Environmental Sciences*, **2(1)**: 70-78.

Montantes, J. (2020). Reasons to use random forest over a neural network: Comparing machine learning versus deep learning. Towards Data Science. Available at: https://medium.com/data-science. Accessed on 16th March, 2026.

Rad, M. R. N., Fanaei, H. R. and Rad, M. R. P. (2015). Application of Artificial Neural Networks to predict the final fruit weight and random forest to select important variables in native population of melon (Cucumis melo L.). *Scientia Horticulturae*, **181**: 108-112.

Raha, S. and Gayen, S. K. (2020). Simulation of meteorological drought using exponential smoothing models: a study on Bankura District, West Bengal, India. *SN Applied Sciences*, **2(5)**: 909.

Rajput, P., Verma, M. K., Ghodichore, N. and Sinha, M. K. (2025). Investigation of spatiotemporal changes of drought trend in Indian River Basin system over the century. *Theoretical and Applied Climatology*, **156(2)**: 100.

Schumann, G., Kirschbaum, D., Anderson, E. and Rashid, K. (2016). Role of earth observation data in disaster response and recovery: From science to capacity building. *In*: *Earth Science Satellite Applications: Current and Future Prospects*. Faisal Hossain (ed.). Springer International Publishing, Cham. pp. 119-146.

Sekulic, S. and Kowalski, B. R. (1992). MARS: a tutorial. *Journal of Chemometrics*, **6(4)**: 199-216.

Shamshad, A., Bawadi, M. A., Hussin, W. W., Majid, T. A. and Sanusi, S. A. M. (2005). First and second order Markov chain models for synthetic generation of wind speed time series. *Energy*, **30(5)**: 693-708.

Shao, Y. E. and Hou, C. D. (2013). Change point determination for a multivariate process using a two-stage hybrid scheme. *Applied Soft Computing*, **13(3)**: 1520-1527.

Shiri, J. (2019). Prediction vs. estimation of dewpoint temperature: assessing GEP, MARS and RF models. *Hydrology Research*, **50(2)**: 633-643.

Sonar, R. B. (2014). Observed trends and variations in rainfall events over Ratnagiri (Maharashtra) during southwest monsoon season. *Mausam*, **65(2)**:171-178.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, **9(3)**: 293-300.

Tettey, M., Oduro, F. T., Adedia, D. and Abaye, D. A. (2017). Markov chain analysis of the rainfall patterns of five geographical locations in the south eastern coast of Ghana. *Earth Perspectives*, *4*(1), 6.

Vapnik, V., Golowich, S. and Smola, A. (1997). Support vector method for function approximation, regression estimation and signal processing. *In: Advances in Neural Information Processing Systems.* M.C. Mozer and M. Jordan and T. Petsche *(eds)*, MIT Press, Cambridge, USA. pp. 281-287.

Velasco, L. C., Aca-ac, J. M., Cajes, J. J., Lactuan, N. J. and Chit, S. C. (2022). Rainfall forecasting using support vector regression machines. *International Journal of Advanced Computer Science and Applications*, **13(3)**: 231-237.

Welling, M. (2004). *Support Vector Regression*. Department of Computer Science, University of Toronto, Toronto (Canada). pp. 3-5

Zhang, G. and Hu, M. Y. (1998). Neural network forecasting of the British pound/US dollar exchange rate. *Omega*, **26(4)**: 495-506.

Sign. of Student	**Sign. of Major Guide**
(P. V. Chovatiya)	**(M. S. Shitap)**